# Nonparametric Statistical Methods for Complete and Censored Data

M.M. Desu
D. Raghavarao

CHAPMAN & HALL/CRC

# Nonparametric Statistical Methods for Complete and Censored Data

Copyrighted Material

Taylor & Francis

Taylor & Francis Group

http://taylorandfrancis.com

Copyrighted Material

# Nonparametric Statistical Methods for Complete and Censored Data

M.M. Desu
D. Raghavarao

CHAPMAN & HALL/CRC

A CRC Press Company
Boca Raton London New York Washington, D.C.

Copyrighted Material

**Library of Congress Cataloging-in-Publication Data**

Desu, M. M.
Nonparametric statistical methods for complete and censored data / M.M. Desu and D. Raghavarao
p. cm.
Includes bibliographical references and index.
ISBN 1-58488-319-7
1. Nonparametric statistics. I. Raghavarao, Damaraju. II. Title.

QA278.8.D47 2003
519.5—dc22 2003060194

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Neither this book nor any part may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, microfilming, and recording, or by any information storage or retrieval system, without prior permission in writing from the publisher.

The consent of CRC Press LLC does not extend to copying for general distribution, for promotion, for creating new works, or for resale. Specific permission must be obtained in writing from CRC Press LLC for such copying.

Direct all inquiries to CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431.

**Trademark Notice:** Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation, without intent to infringe.

**Visit the CRC Press Web site at www.crcpress.com**

© 2004 by Chapman & Hall/CRC

No claim to original U.S. Government works
International Standard Book Number 1-58488-319-7
Library of Congress Card Number 2003060194
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

*Copyrighted Material*

To: Aruna, Subbarao, Anu, Sheila, and Alyssa
M.M.D.

To: Lakshmi,Venkatrayudu, and Sharada
D.R.

Copyrighted Material

Taylor & Francis

Taylor & Francis Group

http://taylorandfrancis.com

Copyrighted Material

# Preface

Nonparametric statistical methods are extremely useful for researchers in biostatistics, pharmaceutical statistics, business, psychology, and social sciences. These methods are precursors for the tools used in analyzing right-censored data. Few books deal extensively with nonparametric statistical methods and pave the way to the analysis of censored data.

This book fills this gap and discusses most of the commonly used nonparametric methods for complete data and then extends those methods to censored data settings. This book can be used as a textbook for a one-semester junior-senior or first-year graduate course. It will also be a useful reference book for researchers who are analyzing censored data or complete data with nonparametric methods.

This is not a theorem–proof format book. While most of the available books are either cookbook type or highly mathematical, this book attempts to introduce the concepts intuitively with minimal mathematical statistics background. Most of the methods discussed are in relation to a univariate response variable. Methods for the analysis of complete data with binary, categorical, and continuous variables are given initially in each setting and then extended to right-censored data on a continuous response. The main text is free of difficult mathematical details, which enables the reader to follow the discussion easily and master the details. The omitted mathematical derivations and other details are given in Appendix A at the end of each chapter. These details can be mastered by individuals with one or two semesters of mathematical statistics training. To facilitate the understanding of the methods, computer programs are given in Appendix B to each chapter. These programs are written in the SAS language so they can be run on the SAS system. The coding for the programs can be found on the CRC Press website, www.crcpress.com, under electronic products/downloads/updates.

In addition to nonparametric methods for analyzing complete and censored data, this book provides excellent discussions on

1. optimal linear rank statistics
2. clinical equivalence
3. analysis of block designs
4. precedence tests

Copyrighted Material

We want to thank Professor Richard N. Schmidt for his continued encouragement and his enormous help in the preparation of the manuscript. We thank our families for their encouragement and continued support.

**M. M. Desu**
**D. Raghavarao**

Copyrighted Material

# Contents



*Copyrighted Material*

Copyrighted Material

Copyrighted Material

Copyrighted Material

Copyrighted Material

Copyrighted Material

CHAPTER 1

# Procedures for a single sample

## 1.1 Introduction

In this chapter we consider procedures for analyzing a random sample on the response variable $X$. Two cases are of interest: (1) $X$ is binary and (2) $X$ is continuous. First we discuss some statistical problems concerning a sample with binary data. Then we discuss procedures for dealing with data on a continuous response variable. We discuss methods for complete data, then methods for censored data situations.

## 1.2 Binary response

A researcher is interested in studying the effectiveness of a new drug under development. For this purpose, suppose 14 patients were recruited and treated. The researcher will be interested in further investigations of the drug if the drug is effective in more than 20% of patients. The researcher may like to know how many of the 14 treated patients should find the drug effective in order that further study is warranted. Furthermore, if 4 of the 14 treated patients found the drug effective, the researcher may like to set up a confidence interval for the probability of effectiveness of the drug. Similarly, a marketing company developed a new commercial and showed it to 30 respondents. Five people liked the commercial. The company wants to set up a confidence interval for the probability of liking this commercial. If 4 out of 30 examinees answered a question incorrectly, does this constitute evidence that 10% of the examinees answered the question incorrectly? Problems of this type also occur in other branches of research and we will discuss these issues in this section.

Consider a random experiment with only two possible outcomes. Traditionally, the outcomes are called *success* and *failure* and the experiment is usually referred to as a *Bernoulli trial*. The probability model for this Bernoulli trial is

$$P(success) = \theta, \quad and \quad P(failure) = 1 - \theta,$$

where $0 < \theta < 1$. In order to learn about $\theta$, the *success probability*, one usually repeats such a Bernoulli trial a fixed number of times, say $n$, where the repetitions are independent. The entire experiment is called a *binomial experiment with n trials*. In relation to each trial we define a random variable. Suppose that $X_i$ is the random variable denoting the outcome of the $i$th trial $(i = 1, 2, \ldots, n)$. The variable $X_i$ takes the value 1, when the outcome is a

Copyrighted Material

"success," and the value 0, otherwise. Thus the probability model for $X_i$ is defined by the probability function

$$f(x;\theta) = P(X_i = x) = \theta^x(1-\theta)^{1-x}, \quad x = 0, 1. \tag{1.1}$$

where $0 < \theta < 1$. The data are the set of observations on the random variables $X_1, X_2, \ldots, X_n$, where these variables are i.i.d. (independent and identically distributed) random variables with the common distribution defined by the probability function $f(x;\theta)$ given in (1.1). This common distribution is called the Bernoulli distribution with the parameter $\theta$ and the data is called a random sample, of size $n$, from a Bernoulli distribution.

The two statistical problems of interest are: (1) the estimation of $\theta$ (point estimation and interval estimation, and (2) testing a hypothesis about the value of $\theta$. The researcher also may be interested in determining $n$, the sample size to meet the objectives of the study.

### *1.2.1 Estimation of success probability*

The point estimate can be obtained from the maximum likelihood method. It is known that the maximum likelihood estimate of $\theta$ is the proportion of successes, i.e.,

$$\hat{\theta} = \Sigma_i X_i/n = S_n/n = X_n. \tag{1.2}$$

It should be noted that the statistic $S_n$ denotes the number of successes.

Binomial distribution

Let $X$ be the number of successes in a binomial experiment with $n$ trials and probability (of success) $\theta$. Then the probability function of $X$ is

$$f(x;n,\theta) = P(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}. \tag{1.3}$$

for $x = 0, 1, \ldots, n$. Here $0 < \theta < 1$.

We denote such a variable $X$ by $Bin(n,\theta)$ and $X$ is said to have the binomial distribution. Sometimes the parameter $n$ is called the *index* and the parameter $\theta$ is called the *probability*. The probability function (1.3) reduces to the probability function (1.1) of the Bernoulli distribution when $n = 1$. In later sections we need to use the cumulative distribution function (cdf) of the binomial distribution and so we note some results about the cdf. For real $x$, the cdf $F$ is

$$F(x;n,\theta) = P(X \leq x).$$

Clearly

$$F(x;n,\theta) = \begin{cases} 0, & \text{for } x < 0, \\ 1, & \text{for } x \geq n. \end{cases}$$

Copyrighted Material

However, for $0 \leq x < n$, we have

$$F(x; n, \theta) = \sum_{i=0}^{j} \binom{n}{i} \theta^i (1-\theta)^{n-i}, \tag{1.4}$$

where $j$ is the integral part of $x$. This sum can be related to an incomplete beta function, which is an integral.

Incomplete beta function

The *incomplete beta function* $I(x; a, b)$ is defined for positive constants $a$ and $b$ and for $0 \leq x \leq 1$ as

$$I(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x u^{a-1}(1-u)^{b-1} du,$$

where $\Gamma(.)$ is the usual gamma function. It may be noted that $I(1; a, b) = 1$, and $I(0; a, b) = 0$. It can be shown that, for $0 \leq x < n$,

$$F(x; n, \theta) = (n-j)\binom{n}{j} \int_0^{1-\theta} u^{n-j-1}(1-u)^j du = I(1-\theta; n-j, j+1). \tag{1.5}$$

where $j$ is the integral part of $x$. The proof concerning the integral representation appears in Appendix A1. From the integral representation (1.5), it is easy to see that the cdf of the binomial distribution is a decreasing function of $\theta$. We also note that

$$E[Bin(n, \theta)] = n\theta, \quad \textit{and} \quad var[Bin(n, \theta)] = n\theta(1-\theta). \tag{1.6}$$

We recall that the statistic $S_n$ follows the binomial distribution with parameters $n$ and $\theta$. Hence from (1.6), it follows that

$$E(\hat{\theta}) = \frac{1}{n} E[Bin(n, \theta)] = \theta. \tag{1.7}$$

So $\hat{\theta}$ is an unbiased estimator of $\theta$. Further,

$$var(\hat{\theta}) = \frac{1}{n^2} var[Bin(n, \theta)] = \theta(1-\theta)/n. \tag{1.8}$$

For the construction of a confidence interval we need an estimate of this variance. An unbiased estimator of this variance is

$$v^2 = [\hat{\theta}(1-\hat{\theta})]/(n-1).$$

In large samples, the distribution of

$$Z = (\hat{\theta} - \theta)/v$$

Copyrighted Material

can be approximated by the standard normal distribution. Using this result, a $100(1-\alpha)\%$ confidence interval for $\theta$ is $(\theta_l, \theta_u)$, where

$$\theta_l = \hat{\theta} - z_{1-\alpha/2} \cdot v, \quad \textit{and} \quad \theta_u = \hat{\theta} + z_{1-\alpha/2} \cdot v, \tag{1.9}$$

with $z_p$ the $100p$ percentile of the standard normal distribution.

A detailed discussion about the confidence intervals is given in Subsection 1.2.7.

### *1.2.2 Testing one-sided hypotheses about $\theta$*

First we consider the problem of testing the simple null hypothesis

$$H_0 : \theta = \theta_0. \tag{1.10}$$

against the simple one-sided alternative hypothesis

$$H_A : \theta = \theta_1 (> \theta_0). \tag{1.11}$$

The Neyman-Pearson lemma can be used to get the most powerful test. This test is to

$$\textit{reject } H_0 \textit{ if } S_n \geq C_+, \tag{1.12}$$

where the constant $C_+$ is chosen so that

$$P(\textit{type I error}) \leq \alpha.$$

In other words, $C_+$ is the smallest integer such that

$$P(S_n \geq C_+ \mid \theta_0) = P(Bin(n, \theta_0) \geq C_+) \leq \alpha. \tag{1.13}$$

In some applications, it is appropriate to use the composite version of (1.11), which is

$$H_+ : \theta > \theta_0. \tag{1.14}$$

For this problem we also use the test (1.12), since the critical value $C_+$ depends only on $\theta_0$, not on $\theta_1$.

The most general problem is concerned with testing the composite null hypothesis

$$H_0^* : \theta \leq \theta_0 \tag{1.15}$$

against the composite (one-sided) alternative hypothesis $H_+$ of (1.14). It turns out that the test (1.12) is also the most powerful test for this general testing problem. This assertion follows from Theorem 8.3.2 of Casella and Berger (1990).

Now let us consider testing the null hypothesis (1.10) against the other one-sided alternative hypothesis,

$$H_- : \theta < \theta_0. \tag{1.16}$$

Copyrighted Material

Table 1.1 *Tests for one-sided alternatives*

| Null Hypothesis | Alternative Hypothesis | Critical Region |
|---|---|---|
| $H_0 : \theta = \theta_0$ | $H_+ : \theta > \theta_0$ | $S_n \geq C_+$ |
| $H_0^* : \theta \leq \theta_0$ | $H_+ : \theta > \theta_0$ | $S_n \geq C_+$ |
| $H_0 : \theta = \theta_0$ | $H_- : \theta < \theta_0$ | $S_n \leq C_-$ |
| $H_0^{**} : \theta \geq \theta_0$ | $H_- : \theta < \theta_0$ | $S_n \leq C_-$ |

We also need to consider the more general problem of testing

$$H_0^{**} : \theta \geq \theta_0 \tag{1.17}$$

against the alternative $H_-$ of (1.16). An analysis similar to the above gives the test. This test is to

$$\textit{reject the null hypothesis if } S_n \leq C_-, \tag{1.18}$$

where $C_-$ is the largest integer such that

$$P(S_n \leq C_- \mid \theta_0) = P(Bin(n, \theta_0) \leq C_-) \leq \alpha. \tag{1.19}$$

A summary of the one-sided tests appears in Table 1.1.

A computer program for obtaining the critical values, $C_+$ and $C_-$, is given in Appendix B1. However, we can approximate the distribution of

$$Z(\theta) = \frac{S_n - n\theta}{\sqrt{n\theta(1-\theta)}} \tag{1.20}$$

by the standard normal distribution, when $min\{n\theta, n(1-\theta)\} \geq 5$. Using this result, we obtain approximations to the critical values $C_+$ and $C_-$. Starting from equation (1.13), and using the continuity correction, we have the condition

$$P(S_n \geq C_+ - 0.5 \mid \theta_0) \leq \alpha.$$

In turn, this condition is the same as

$$P\left(\frac{S_n - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} \geq C_+^*\right) \leq \alpha,$$

where

$$C_+^* = \frac{C_+ - 0.5 - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}.$$

The condition on the probability can be restated as

$$P(Z(\theta_0) \leq C_+^*) \geq 1 - \alpha.$$

Under $H_0$, a normal approximation can be used for the distribution of the statistic $Z(\theta_0)$. So we can satisfy the above condition by choosing $C_+^*$ as

Copyrighted Material

$z_{(1-\alpha)}$, the $100(1-\alpha)$ percentile of the standard normal distribution. Thus an approximation to $C'_+$ is

$$C'_+ \approx n\theta_0 + 0.5 + z_{(1-\alpha)}\sqrt{n\theta_0(1-\theta_0)}.$$

Since we want an integer value for $C'_+$, this approximation is restated as

$$C_+ \approx \lfloor n\theta_0 + 0.5 + z_{(1-\alpha)}\sqrt{n\theta_0(1-\theta_0)} \rfloor + 1. \tag{1.21}$$

where $\lfloor x \rfloor$ denotes the integral part of $x$.

A similar analysis gives

$$C'_- \approx \lfloor n\theta_0 - 0.5 + z_\alpha \sqrt{n\theta_0(1-\theta_0)} \rfloor. \tag{1.22}$$

Equations (1.21) and (1.22) give very good approximations, whenever $min\{n\theta_0, n(1-\theta_0\} \geq 5$. For example, when $n = 20$, $\theta = 0.25$, and $\alpha = 0.05$, the exact values are obtained using the computer program given in Appendix B1. These are $C_+ = 9$, and $C_- = 1$. From equations (1.21) and (1.22) the approximations are $C_+ \approx 9$ and $C'_- \approx 1$. In this case the approximations are the same as the exact values.

### *1.2.3 P-values for one-sided tests*

Instead of calculating the critical values and performing the test, one can compute the $P$-value (of the data), which is a measure of the strength of evidence against the null hypothesis, and compare it with the chosen $\alpha$ value. Let $s$ be the observed value of the statistic $S_n$. The $P$-value for the test (1.12) is

$$P_+ = P(Bin(n, \theta_0) \geq s), \tag{1.23}$$

and when $min\{n\theta_0, n(1-\theta_0\} \geq 5$, an approximation is

$$P_+ \approx \Phi[(n\theta_0 - s + 0.5)/\sqrt{n(\theta_0(1-\theta_0)}]. \tag{1.24}$$

The $P$-value for the test (1.18) is

$$P_- = P(Bin(n, \theta_0) \leq s), \tag{1.25}$$

and when $min\{n\theta_0, n(1-\theta_0\} \geq 5$, an approximation is

$$P_- \approx \Phi[(s + 0.5 - n\theta_0)/\sqrt{n(\theta_0(1-\theta_0)}]. \tag{1.26}$$

In (1.24) and (1.26), $\Phi(.)$ is the cdf of the standard normal distribution. It is customary to give the $P$-value while reporting the results, and the computer programs usually report the $P$-values for tests.

We can also use the $P$-value for performing a test of hypothesis, as mentioned earlier. This method can be stated as follows:

*Reject the null hypothesis if* $P$-value $\leq \alpha$. (1.27)

Copyrighted Material

Example 1.1. The first example discussed at the beginning of Section 1.2 can be formulated as a problem of testing the null hypothesis

$$H_0^* : \theta \leq 0.2 \textit{ against the alternative } H_+ : \theta > 0.2.$$

Let us take $\alpha = 0.05$. From the computer program we get $C_+ = 6$. Thus under the test (1.12), the researcher should reject $H_0^*$ in favor of $H_+$ and develop the drug further when $S_{14} \geq 6$.

Suppose as indicated before that $S_{14} = 4$. The exact $P$-value from (1.23) is

$$P_+ = P(Bin(14, 0.2) \geq 4) = 0.3017,$$

which is obtained from the corresponding SAS function. Since $P_+ > \alpha = 0.05$, we do not reject $H_0^* : \theta \leq 0.2$.

Now we will study the power function of the test (1.12). It will be used for designing a study, which is the subject of Subsection 1.2.5.

### *1.2.4 Power function of one-sided tests*

The power function of the test (1.12) is

$$\begin{aligned}
\pi_+(\theta) &= P(\textit{rejecting } H_0 \mid \theta) \\
&= P(S_n \geq C_+ \mid \theta) \\
&= 1 - P(S_n \leq C_+ - 1 \mid \theta) \\
&= 1 - F(C_+ - 1; n, \theta) \\
&= 1 - I(1 - \theta; n - C_+ + 1, C_+) \\
&= I(\theta; C_+, n - C_+ + 1)
\end{aligned}$$

The last equality follows from the previous one by changing the variable of integration (see Appendix A1). From the integral representation, it is easy to see that this power function is an increasing function of $\theta$. An approximation to the power function is useful for determining the size of an experiment. Using the normal approximation to the binomial distribution, an approximation to the power function is derived. Since the power function of test (1.12) is

$$\pi_+(\theta) = 1 - P(S_n \leq C_+ - 1 \mid \theta),$$

the normal approximation for the distribution of $S_n$, with continuity correction, gives the approximation

$$\pi_+(\theta) \approx 1 - \Phi[(C_+ - 0.5 - n\theta)/\sqrt{n\theta(1-\theta)}].$$

Using the symmetry property of the normal cdf, we can simplify the right-hand-side expression and then we have

$$\pi_+(\theta) \approx \Phi[(n\theta + 0.5 - C_+)/\sqrt{n\theta(1-\theta)}]. \tag{1.28}$$

Similarly, the power function of the test (1.18) can be seen to be

$$\pi_-(\theta) = I(1 - \theta; n - C_-, C_- + 1), \tag{1.29}$$

Copyrighted Material

and it can be approximated as

$$\pi_-(\theta) \approx \Phi[(C_- + 0.5 - n\theta)/\sqrt{n\theta(1-\theta)}]. \tag{1.30}$$

*1.2.5 Sample size*

We want to set up a study with $n$ trials to test $H_0 : \theta = \theta_0$ versus $H_+ : \theta > \theta_0$, at a significance level $\alpha$. Consequently, the problem is to decide upon the sample size $n$ so that the test, based on our study, has adequate power for all $\theta \geq \theta_1 (> \theta_0)$. We want the power of the test (1.12) to be at least $1 - \beta$ for all $\theta \geq \theta_1$. In view of the monotone property of the power function, this requirement on the power is satisfied by requiring the power at $\theta_1$ to be at least $1 - \beta$. Here, for convenience, we denote the critical value by $c$. Using the power function expression, the requirements are

$$\pi_+(\theta_0) \leq \alpha\,;\ \pi_+(\theta_1) \geq 1 - \beta.$$

These requirements are the same as

$$I(\theta_0; c, n - c + 1) \leq \alpha\,;\ I(\theta_1; c, n - c + 1) \geq 1 - \beta. \tag{1.31}$$

In the power function expression we had the constant $C_+$ and this is replaced by $c$ for convenience. Thus we need to choose $n$ and $c$ so as to satisfy the inequalities (1.31).

An iterative technique is needed to find the required $n$ and $c$. To start the iteration one can use approximations for $n$ and $c$. Now we obtain a set of useful approximations.

Using the normal approximation with the continuity correction for the power function and changing the inequalities to equalities in (1.31), two equations are obtained. These are

$$\Phi\left[\frac{n\theta_0 + 0.5 - c}{\sqrt{n\theta_0(1-\theta_0)}}\right] = \alpha; \Phi\left[\frac{n\theta_1 + 0.5 - c}{\sqrt{n\theta_1(1-\theta_1)}}\right] = 1 - \beta.$$

These equations are the same as

$$\frac{n\theta_0 + 0.5 - c}{\sqrt{n\theta_0(1-\theta_0)}} = z_\alpha;\ \frac{n\theta_1 + 0.5 - c}{\sqrt{n\theta_1(1-\theta_1)}} = z_{1-\beta}.$$

Solving these equations, approximations for the required sample size and the critical value are obtained. The solution is $(n^*, c^*)$, where

$$n^* = \frac{\left[z_\alpha\sqrt{\theta_0(1-\theta_0)} - z_{1-\beta}\sqrt{\theta_1(1-\theta_1)}\right]^2}{(\theta_1 - \theta_0)^2}$$

$$= \frac{\left[z_\alpha\sqrt{\theta_0(1-\theta_0)} + z_{(\beta)}\sqrt{\theta_1(1-\theta_1)}\right]^2}{(\theta_1 - \theta_0)^2}.$$

Copyrighted Material

We can also rewrite the formula as

$$n^* = \frac{\left[z_{1-\alpha}\sqrt{\theta_0(1-\theta_0)} + z_{1-\beta}\sqrt{\theta_1(1-\theta_1)}\right]^2}{(\theta_1-\theta_0)^2} \equiv \frac{A}{(\theta_1-\theta_0)^2}, \tag{1.32}$$

and

$$c^* = n^*\theta_0 + 0.5 - z_\alpha\sqrt{n^*\theta_0(1-\theta_0)}. \tag{1.33}$$

Thus an integer approximation to the sample size is

$$n_u \approx \lfloor n^* \rfloor + 1. \tag{1.34}$$

An integer approximation to $c$ is

$$c_u \approx \lfloor c^* \rfloor + 1. \tag{1.35}$$

A better approximation can be obtained using the results of Levin and Chen (1999) and these modified values will be given now. The $n^*$ is modified as

$$n_L = \frac{n^*}{4}[1 + \sqrt{1 + 2(\theta_1 - \theta_0)/A}]^2,$$

where $A$ is defined in (1.32) and the $c^*$ is modified as

$$c_L = n_L\theta_0 + 0.5 - z_\alpha\sqrt{n_L\theta_0(1-\theta_0)}.$$

Using these values the modified integer approximations are

$$n_{LC} \approx \lfloor n_L \rfloor + 1 \tag{1.36}$$

and

$$c_{LC} \approx \lfloor c_L \rfloor + 1. \tag{1.37}$$

A computer program for doing these calculations is given in Appendix B1.

This sample size problem is the same as the problem of designing a *Phase II clinical trial* as discussed by Thall and Simon (1995). They give a table of $n$ and $c$ values that are solutions to (1.31). In connection with the Phase II trials, Thall and Simon indicate that reasonable values for the difference $(\theta_1 - \theta_0)$ are from 0.15 to 0.20. Now we illustrate the calculation of approximations (1.34) and (1.35) with an example.

Example 1.2. While testing $H_0^* : \theta \leq 0.2$ against $H_+ : \theta > 0.2$ with $\alpha = 0.05$, the experimenter wants to have a power of 0.80 for the test, when $\theta = 0.35$. Then from (1.32), we have

$$n^* = [(1.645\sqrt{(0.2)(0.8)} + 0.84\sqrt{(0.35)(0.65)}]^2/(0.35-0.2)^2] = 49.81.$$

Using this value in (1.34), we get

$$n_u \approx \lfloor 49.81 \rfloor + 1 = 50.$$

Copyrighted Material

Using the $n^*$ value in (1.33) we get the $c^*$ value and using this $c^*$ in equation (1.35) we get $c_u \approx 16$. Thus the rejection region of an approximate 0.05-level test is $S_{50} \geq 15$. The modified approximations will turn out to be $n_{LC} = 57$, and $c_{LC} = 17$. These are taken from the output of a computer program given in Appendix B1. With the modified solution, the error probabilities are much closer to the specifications, compared to the unmodified solution.

Another approximation to the sample size can be obtained by using the arcsine transform of the statistic $\sqrt{(S_n/n)}$. The relevant details are given in Desu and Raghavarao (1990). This derivation is assigned as Problem 2. Using this transformation, Natrella (1963) prepared a table of the $n$-values.

### *1.2.6 Testing a two-sided hypothesis about $\theta$*

Suppose we want to test the simple null hypothesis (1.10) that $\theta = \theta_0$ against the two-sided composite alternative hypothesis

$$H_A : \theta \neq \theta_0. \tag{1.38}$$

The usual test is to

$$\textit{reject } H_0 \textit{ if } S_n \leq c_1 \quad \textit{or} \quad S_n \geq c_2, \tag{1.39}$$

where $c_1$ is the largest integer and $c_2$ is the smallest integer such that

$$P(S_n \leq c_1 \mid H_0) \leq (\alpha/2);\ P(S_n \geq c_2 \mid H_0) \leq (\alpha/2). \tag{1.40}$$

This test can be derived from the *union-intersection principle*. The details of this derivation appear in Appendix A1. Using the normal approximation for the binomial distribution we can obtain approximations for the required $c_1$ and $c_2$ values. In particular, these approximations are

$$c_1 \approx \lfloor n\theta_0 - 0.5 + z_{(\alpha/2)}\sqrt{n\theta_0(1-\theta_0)} \rfloor,$$

and

$$c_2 \approx \lfloor n\theta_0 + 0.5 + z_{(1-\alpha/2)}\sqrt{n\theta_0(1-\theta_0)} \rfloor + 1. \tag{1.41}$$

An important special case is the one for which $\theta_0 = 0.5$, and it will be discussed in Subsection 1.3.2.

Remark 1.1. In this case, the $P$-value is usually computed as $2\,min(P_+, P_-)$, where $P_+$ and $P_-$ are given by (1.23) and (1.25).

### *1.2.7 Confidence intervals for $\theta$*

In some applications the researcher may be interested in a confidence interval for $\theta$. We need to find two functions $\theta_L(S_n)$ and $\theta_U(S_n)$ of the random variable $S_n$ such that

$$P(\theta_L(S_n) < \theta < \theta_U(S_n)) \geq 1 - \alpha. \tag{1.42}$$

Copyrighted Material

Then the interval $(\theta_L(S_n), \theta_U(S_n))$ is a confidence interval for $\theta$ with confidence coefficient $(1-\alpha)$. These limits are usually derived from the acceptance region of the two-sided test (1.39). As a prelude to this derivation, we consider the problem of finding one-sided confidence limits or confidence bounds. From these bounds we can get a confidence interval.

An *upper confidence bound* $\theta_{UB}(S_n)$ is a function of $S_n$ such that

$$P(\theta < \theta_{UB}(S_n)) \geq 1-\alpha, \tag{1.43}$$

and a *lower confidence bound* $\theta_{LB}(S_n)$ is a function of $S_n$ such that

$$P(\theta_{LB}(S_n) < \theta) \geq 1-\alpha. \tag{1.44}$$

From these bounds we get the one-sided confidence intervals $(\theta_{LB}(S_n), 1)$ and $(0, \theta_{UB}(S_n))$. In the case of Phase II trials one wants to ensure that the response rate is not too low. A lower bound for the response rate will enable a researcher to decide to proceed or not with the development of a new drug. An upper bound for the proportion of nonconforming units will enable an engineer to accept or reject manufactured items supplied by a vendor.

Upper confidence bound

To derive an upper confidence bound consider the lower tail $\alpha$-level test (1.18). Let $s$ be the observed value of $S_n$. Under this test we reject the null $H_0 : \theta = \theta_0$, if $s \leq c$, where $P(S_n \leq c|\theta_0) \leq \alpha$. In other words, we reject $H_0$ if

$$F(s; n, \theta_0) \leq \alpha,$$

where $F$ is the cdf of $S_n$. Since the cdf is a decreasing function of $\theta_0$, we can find $\theta_{UB}$ such that

$$F(s; n, \theta_{UB}) = \alpha.$$

Note that $\theta_{UB}$ is a function of $s$. We also have

$$F(s; n, \theta_0) \leq \alpha, \quad \textit{for } \theta_0 \geq \theta_{UB},$$

and

$$F(s; n, \theta_0) > \alpha, \quad \textit{for } \theta_0 < \theta_{UB}.$$

Thus we do not reject $H_0$ for $\theta_0 < \theta_{UB}(s)$, where $s$ is the observed value of $S_n$. As the probability of not rejecting is at least $1-\alpha$, we have

$$P(\theta < \theta_{UB}(S_n)|\theta) \geq 1-\alpha.$$

Thus $\theta_{UB}(S_n)$ is a $(1-\alpha)$ upper confidence bound for $\theta$. Using the incomplete beta function representation for the cdf of $S_n$, the bound $\theta_{UB}$ can be seen as the solution of the equation

$$I(\theta_{UB}; S_n + 1, n - S_n) = 1-\alpha.$$

Copyrighted Material

This leads to the formula

$$\theta_{UB}(S_n) = BINV(1-\alpha; S_n + 1, n - S_n) \tag{1.45}$$

for $S_n < n$, where $BINV(p; a, b)$ is the $100p$ percentile of the $Beta(a, b)$ distribution. The beta percentiles are standard SAS functions. When $S_n = n$, the upper bound is taken as one.

Lower confidence bound

A similar analysis will give an expression for the *lower confidence bound* as

$$\theta_{LB}(S_n) = BINV(\alpha; S_n, n - S_n + 1) \tag{1.46}$$

for $S_n > 0$ and for $S_n = 0$ the lower bound is taken as zero.

For example, suppose an inspector examined a sample of 100 items and found that 3 of them are defective. To decide whether or not to accept the lot, an upper bound is calculated. For these data the 95% upper bound for $\theta$ is

$$\theta_{UB}(3) = BINV(0.95; 4, 97),$$

which turns out to be 0.0757, that is, 7.57%. So, in the worst case scenario, the percentage of nonconforming units could be as high as 7.57%. If this percentage is larger than the acceptable percentage, the lot would be rejected.

For example, suppose $S_{14}$ is 4. Then the lower bound is

$$\theta_{LB}(4) = BINV(0.05; 4, 11) = 0.104.$$

The researcher will proceed further only if there is evidence that $\theta$ is at least 0.2. Because this lower bound is less than 0.2, further development of the drug will not be pursued.

Exact confidence limits

Let $\theta_L(S_n)$ be the lower $(1 - \alpha/2)$ confidence bound and $\theta_U(S_n)$ be the upper $(1 - \alpha/2)$ confidence bound. Then we have

$$P[\theta_L(S_n) < \theta < \theta_U(S_n)] = 1 - P(\theta \leq \theta_L(S_n)) - P(\theta_U(S_n) \geq \theta).$$

However,

$$P(\theta_L(S_n) < \theta) \geq (1 - \alpha/2) \Rightarrow -P(\theta \leq \theta_L(S_n)) \geq -(\alpha/2)$$

and

$$P(\theta < \theta_U(S_n)) \geq (1 - \alpha/2) \Rightarrow -P(\theta_U(S_n) \geq \theta) \geq -(\alpha/2).$$

Hence

$$P[\theta_L(S_n) < \theta < \theta_U(S_n)] \geq 1 - (\alpha/2) - (\alpha/2) = 1 - \alpha.$$

Copyrighted Material

In other words $(\theta_L(S_n), \theta_U(S_n))$ is a $(1-\alpha)$ confidence interval for $\theta$. Thus, for $0 < S_n < n$ these confidence limits are given by

$$\theta_L(S_n) = BINV(\alpha/2; S_n, n - S_n + 1);$$

and

$$\theta_U(S_n) = BINV(1 - \alpha/2; S_n + 1, n - S_n). \tag{1.47}$$

If $S_n = 0$, the lower limit is taken as zero and if $S_n = n$, the upper limit is taken as 1. Several tabulations of these limits have been made. One reference is the set of tables edited by Lentner (1982).

Example 1.3. Suppose we observed 3 successes in 20 trials. We want to find a 95% confidence interval for the parameter $\theta$, the success probability. The point estimate $\hat{\theta} = (3/20) = 0.15$. Using the formula (1.47) and a SAS program (see Appendix B1), we get the confidence limits $\theta_L = 0.0321$ and $\theta_U = 0.3789$. In other words, a 95% confidence interval for the parameter $\theta$ is $(0.0321, 0.3789)$.

Confidence limits using the asymptotic distribution

For large or moderate $n$, the confidence limits are usually derived using the normal approximation to the distribution of $\hat{\theta}$. In elementary textbooks the interval

$$(\hat{\theta} - z_{(1-\alpha/2)} \cdot v_1,\ \hat{\theta} + z_{(1-\alpha/2)} \cdot v_1) \tag{1.48}$$

is suggested as a confidence interval where $\hat{\theta}$ is given by (1.2) and $v_1^2 = [\hat{\theta}(1-\hat{\theta})]/n$, a biased estimator of $var(\hat{\theta})$.

Samuels and Lu (1992) give a set of guidelines for deciding the situations when this interval provides a good answer.

Ghosh (1979) has investigated and recommended a method that is as simple as the above method for constructing a confidence interval and as good as the exact method. We give the result here and the method of derivation is relegated to Problem 1. This confidence interval for $\theta$ is

$$((\hat{\theta} + C - z_{(1-\alpha/2)} \cdot v_*)/(1+2C), (\hat{\theta} + C + z_{(1-\alpha/2)} \cdot v_*)/(1+2C)), \tag{1.49}$$

where

$$C = (z_{(1-\alpha/2)})^2/2n;\ v_*^2 = [\hat{\theta}(1-\hat{\theta}) + (C/2)]/n = v_1^2 + (C/2n).$$

Recent studies of Agresti and Coull (1998) and Newcombe (1998) reinforced the recommendation of the interval (1.49). Also see Agresti and Caffo (2000) for further discussion on the confidence interval of $\theta$. A computer program for calculating the interval (1.49) is given in Appendix B1.

For example, for $n = 20$ and $s = 3$ the 95% confidence interval (1.49) is (0.0523, 0.3604).

Copyrighted Material

Using a confidence interval for testing

A confidence interval can be used to test the simple null hypothesis (1.10) that $\theta = \theta_0$ against the two-sided alternative (1.38). The corresponding test is to

$$\textit{reject } H_0 \textit{ if } \theta_0 \textit{ is not in the interval.} \tag{1.50}$$

## 1.3 Complete data on continuous responses

In some studies the response variable can be viewed as a continuous random variable. In reliability studies and in clinical trials the response variable is *time to an event*. It may be the time to first breakdown of a machine or time to death of a patient with terminal cancer. In these studies we want to estimate some characteristics of the distribution of the variable of interest. Suppose we are interested in studying the properties of lifetime distributions.

Dunsmore (1974) obtained data on time to first breakdown for 20 machines. This set of 20 machines is viewed as a random sample. The time to first breakdown is the response variable. The data obtained here are observations on i.i.d. random variables $X_1, X_2, \ldots, X_n$, where the common probability distribution is defined by some probability density function $f(x)$. We have very limited knowledge about the density function. The objective is to estimate some characteristics of the (population) distribution of the time to first breakdown.

We assume that the probability distribution has a unique median. We want to estimate this median, which is usually used as a measure of location. In some cases we may want to test a hypothesis about the median. For example, a social scientist may be interested in testing that the median annual family income in a county is \$25,000 based on a random sample of family annual income data. This testing problem is also of interest in the evaluation of a cancer treatment, where the efficacy of a treatment is characterized by the median survival time. In clinical studies, observations on some subjects frequently are not complete, since different subjects enter the study at different times and for some subjects the event did not occur before the end of the study. These incomplete observations are called *right-censored observations*. In this section we discuss the results for complete data situations. Some generalizations for the censored data cases are discussed in Section 1.4.

### *1.3.1 Point estimation of the median*

We have observations on $(X_1, X_2, \ldots, X_n)$, a random sample from the distribution defined by the pdf $f(.)$. The cdf of the population distribution is $F(.)$. Using these data we want to estimate the median, $\xi$, and test a hypothesis about the median.

The intuitive choice for the point estimator is the sample median. To give an expression for the sample median we need the order statistics of the sample. These are the sample values arranged in increasing order of magnitude.

Copyrighted Material

We denote these order statistics by $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, where $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$. A point estimate of $\xi$ is $\hat{\xi}$, the sample median, and is defined as

$$\hat{\xi} = \begin{cases} X_{(k+1)}, & \text{if } n = 2k+1, \\ (X_{(k)} + X_{(k+1)})/2, & \text{if } n = 2k. \end{cases}$$

The sampling distribution of this statistic depends on the population distribution in a complicated way. However, some properties of this estimator can be obtained by making certain assumptions about the population density function. Desu and Rodine (1969) showed that for symmetric densities, the sample median is an unbiased estimator of the population median, which is equal to the population mean. The interested reader is referred to their paper for the proofs and other details.

Sometimes one order statistic $X_{(s)}$ is used as an estimator of the median, where $s$ is the integer $\lfloor (n/2) \rfloor + 1$. Further discussion along these lines appears in Subsection 1.3.6.

We first discuss the testing problem and then proceed to the problem of finding a confidence interval for the median $\xi$. This discussion can be carried out without any restrictions on the population distribution.

### *1.3.2 Sign test for testing a simple null hypothesis about the median*

Let $\xi$ be the population median. Consider the case of testing the null hypothesis

$$H_0 : \xi = \xi_0 \tag{1.51}$$

against the one-sided alternative

$$H_{A1} : \xi > \xi_0. \tag{1.52}$$

From the definition of the population median it is clear that $P(X_i > \xi) = 1/2$ or $P(X_i - \xi > 0) = 1/2$. Let $\theta = P(X_i - \xi_0 > 0)$. It follows that $\theta = 1/2$ or $> 1/2$ depending on whether (1.51) or (1.52) is true. Thus this hypothesis testing problem can be translated into a testing problem in relation to a binary data set. This translation will be explained now. We transform the data by defining

$$Z_i = \begin{cases} 1, & \text{if } X_i - \xi_0 > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Denoting the $P(Z_i = 1)$ by $\theta$, and using $Z$s, the statistical problem can be restated as that of testing the null hypothesis

$$H_0 : \theta = 1/2, \text{ against the alternative } H_+ : \theta > 1/2. \tag{1.53}$$

From the discussion in Subsection 1.2.2, it is obvious that we can use the test defined by the critical region (1.12). Here the test statistic $S_n$ is equal

Copyrighted Material

to $\sum_i Z_i$. In other words, $S_n$ stands for the number of $X$ values that are greater than $\xi_0$. The critical region of the test is

$$S_n \geq C_+.$$

where the constant $C_+$ is the smallest integer such that

$$P(Bin(n, 1/2) \geq C_+) \leq \alpha. \tag{1.54}$$

For $n \geq 10$, we can approximate $C_+$ of (1.54) as

$$C_+ \approx \lfloor (n/2) + 0.5 + z_{1-\alpha}\sqrt{(n/4)} \rfloor + 1. \tag{1.55}$$

Now let us consider the problem of testing the null hypothesis (1.51) against the other one-sided alternative.

$$H_{A2} : \xi < \xi_0. \tag{1.56}$$

This problem is equivalent to testing the null hypothesis of (1.53) against the other one-sided alternative,

$$H_- : \theta < 1/2. \tag{1.57}$$

Clearly this testing problem can be handled by the test defined by the critical region (1.18).

Suppose the alternative hypothesis is a two-sided one, namely,

$$H_A : \xi \neq \xi_0. \tag{1.58}$$

This testing problem is equivalent to testing

$$H_0 : \theta = 1/2 \textit{ against the alternative } H_A : \theta \neq 1/2. \tag{1.59}$$

The relevant test for this two-sided alternatives case is to

$$\textit{reject } H_0 \textit{ if } S_n \leq C, \quad \textit{or} \quad S_n \geq n - C, \tag{1.60}$$

because the null distribution of $S_n$ is symmetrical under $H_0$. Here $C$ is the largest integer such that

$$P(Bin(n, 1/2) \leq C) \leq (\alpha/2). \tag{1.61}$$

Using the table from MacKinnon (1964), we can obtain this $C$ value. For $n \geq 10$, using the normal distribution approximation to the distribution of $S_n$, $C$ can be approximated as

$$C \approx \lfloor (n/2) - 0.5 + z_{(\alpha/2)}\sqrt{(n/4)} \rfloor. \tag{1.62}$$

Example 1.4. For $n = 10$ and $\alpha = 0.05$, from (1.62) we have

$$C \approx \lfloor 5 - 0.5 - 1.96\sqrt{(2.5)} \rfloor = 1.$$

Using the computer program given in Appendix B1, we find that $C = 1$. Here the normal approximation and the exact value for $C$ coincide.

Copyrighted Material

These procedures can be adopted easily to test hypotheses about other percentiles. Since the $Z$ is 1 or 0 depending on whether the difference $(X - \xi)$ is positive or negative, these tests are sometimes referred to as *sign tests*.

A distribution-free confidence interval for the median $\xi$

This interval can be obtained from the acceptance region of the two-sided test (1.60). Assume that for the given sample size $n$ and the confidence coefficient $(1 - \alpha)$, the constant $C(> 0)$ satisfying (1.61) exists. Let $d = C + 1$. Then the acceptance region of the two-sided test (1.60) can be seen to be

$$\{\xi_0 : d \leq \Sigma_i Z_i \leq n - d\}. \tag{1.63}$$

This means that the number of $X$-values greater than $\xi_0$ is at least $d$ and not more than $n - d$. Thus a 100 $(1 - \alpha)\%$ confidence interval for $\xi$ is $[X_{(d)}, X_{(n-d+1)})$, where $X_{(i)}$'s are the order statistics of the sample. Van der Parren (1970) published a table of $d$-values, which can be used for constructing the confidence intervals. This table also gives the exact coverage probability, which is not available in the table of MacKinnon (1964).

Remark 1.2. In this discussion it is implicitly assumed that there are no ties. When there are tied values in the sample, a modification of this procedure is needed. This modification is given in Subsection 1.3.6.

Example 1.5. Dunsmore (1974) observed 20 machines and obtained data on times (in hours) to first breakdown. We consider only 10 observations. These are

18, 23, 29, 409, 24, 74, 13, 62, 46, and 4.

The order statistics, $X_{(i)}$, of the sample can be seen to be

4, 13, 18, 23, 24, 29, 46, 62, 74, and 409.

The sample median is $[X_{(5)} + X_{(6)}]/2 = 26.5$, which is a point estimate of $\xi$, the population median. In addition, we want a 95% confidence interval for the median $\xi$. Here $n = 10$, and $\alpha = 0.05$. In Example 1.4, we found that $C = 1$ and hence $d = 2$. Thus the confidence interval is $[X_{(2)}, X_{(9)})$. Hence a 95% confidence interval for the median $\xi$ is [13, 74).

### *1.3.3 Estimation of the cdf*

Sample distribution function plays an important role in the analysis of continuous response data. It can be used to obtain estimates of certain probabilities of interest and from it we can also obtain a confidence band for the population distribution function.

Copyrighted Material

Suppose that our sample is $(X_1, X_2, \ldots, X_n)$. The sample distribution function (or empirical distribution function) denoted by $F_n(x)$ is defined as

$$F_n(x) = \{number\ of\ X\ values\ which\ are \leq x\}/n$$
$$= \sum_{h=1}^{n} u(X_h, x)/n, \tag{1.64}$$

where $u(a, b) = 1$, if $a \leq b$ and $= 0$, otherwise. It may be noted that this function depends on the sample values; however, our notation does not indicate this fact.

In general, if $X$ is our response variable, the probability $P(X \leq x) = F(x)$ is estimated by $F_n(x)$, for each real $x$. In other words, for each real $x$,

$$\hat{F}(x) = F_n(x). \tag{1.65}$$

Some properties of this estimator are noted for future use. For a fixed $x$, the statistic $nF_n(x)$ follows a *binomial distribution* with parameters $n$ and $F(x)$. So it follows that

$$E(F_n(x)) = E(nF_n(x))/n = F(x), \quad var(F_n(x)) = F(x)(1 - F(x))/n.$$

By identifying $F(x)$ as $\theta$, $nF_n(x)$ as $S_n$, and $F_n(x)$ as $\hat{\theta}$ in relation to the binary data setting of Section 1.2, we can find an exact or asymptotic confidence interval for $F(x)$. Let $v^2(x)$ be the unbiased estimator of the $var(F_n(x))$, so that

$$v^2(x) = F_n(x)(1 - F_n(x))/(n - 1). \tag{1.66}$$

For large $n$, the distribution of $F_n(x)$ can be approximated by a normal distribution and using this approximate distribution, it can be seen that

$$(\, F_n(x) - z_{(1-\alpha/2)} \cdot v(x), F_n(x) + z_{(1-\alpha/2)} \cdot v(x)\,) \tag{1.67}$$

is a confidence interval for $F(x)$ and the associated confidence coefficient is approximately equal to $(1 - \alpha)$.

Example 1.5 (cont'd.). From the Dunsmore data of Example 1.5, suppose we want to estimate the probability that the time to first breakdown is not greater than 46 hours. This probability is

$$P(X \leq 46) = F(46).$$

So it can be estimated by $F_n(46) = (7/10) = 0.7$. Now let us compute a confidence interval for $F(46)$. We first compute

$$v^2(46) = (7/10)(3/10)/9 = 0.0233,$$

and then

$$z_{0.975} \cdot v(46) = 1.96(.1527) = 0.2994.$$

Copyrighted Material

Using the formula (1.67), we get a confidence interval for $F(46)$ as $(0.7 - 0.2994, 0.7 + 0.2994)$. In other words an approximate 95% confidence interval for $F(46)$ is $(0.4006, 0.9994)$.

### *1.3.4 Estimation of survival function*

In reliability or survival studies, the researcher is interested in estimating the probability of surviving beyond $x$. This probability is

$$S(x) = P(X > x) = 1 - F(x), \tag{1.68}$$

and this function is called the *survival function*. A natural estimator of $S(x)$ is

$$\hat{S}(x) = 1 - \hat{F}(x) = 1 - F_n(x) \equiv S_n(x). \tag{1.69}$$

It is easy to verify that $S_n(x)$ in the above equation is the proportion of $x$ values that are greater than $x$. This function is called the *sample survival function*.

Let us examine this estimator of $S(x)$ in more detail so that we can generalize this result for censored data. Let $Y_1 < Y_2 \ldots < Y_r$ be the distinct ordered values of the random sample of size $n$ and let $d_i$ be the number of times $Y_i$ occurs in the sample. Recursively define $n_1 = n$ and $n_i = n_{i-1} - d_{i-1}$, for $i = 2, 3, \ldots r$. Note that $n_i$ are the number of observations $\geq Y_i$. From (1.69), we have

$$\hat{S}(x) = \begin{cases} 1. & \text{for } x < Y_1, \\ 1 - \frac{\sum_{i=1}^{j} d_i}{n}, & \text{for } Y_j \leq x < Y_{j+1}, j = 1, 2, \ldots, r-1. \\ 0. & \text{for } x \geq Y_r. \end{cases} \tag{1.70}$$

Noting that

$$1 - \frac{d_1 + d_2}{n} = \left(1 - \frac{d_1}{n_1}\right)\left(1 - \frac{d_2}{n_2}\right),$$

we get

$$1 - \frac{\sum_{i=1}^{j} d_i}{n} = \prod_{i=1}^{j} \left(1 - \frac{d_i}{n_i}\right).$$

Thus the expression (1.70) can be rewritten as

$$\hat{S}(x) = \begin{cases} 1. & \text{for } x < Y_1, \\ \Pi_{\{i: Y_i \leq x\}} (1 - \frac{d_i}{n_i}), & \text{for } x \geq Y_1. \end{cases} \tag{1.71}$$

This equation means that the estimated survival probability is the product of the probabilities of surviving in the $Y$-intervals preceding $x$.

It is easy to see that for fixed $x$,

$$E(\hat{S}(x)) = 1 - E(F_n(x)) = 1 - F(x) = S(x),$$

Copyrighted Material

and

$$var(\hat{S}(x)) = var(F_n(x)) = F(x)[1 - F(x)]/n = S(x)(1 - S(x))/n. \quad (1.72)$$

For large $n$, a confidence interval for $S(x)$ with confidence coefficient $1 - \alpha$ is

$$\left(S_n(x) - z_{(1-\alpha/2)} \cdot v(x), S_n(x) + z_{(1-\alpha/2)} \cdot v(x)\right), \quad (1.73)$$

where $v^2(x)$ is given by (1.67).

Example 1.6. For the Dunsmore data of Example 1.5, suppose we want to estimate the probability that the first breakdown occurs after 46 hours. This probability is $S(46) = 1 - F(46)$. Thus $S_n(46) = 1 - F_n(46) = 0.3$. It is easy to see that an approximate 95% confidence interval for $S(46)$ is $(0.3 - 0.2994, 0.3 + 0.2994)$. In other words, the required confidence interval is (0.0006, 0.5994).

Remark 1.3. Since $F(x)$ and $S(x)$ are probabilities we can use the exact methods of Subsection 1.2.7 for constructing the confidence intervals. Here we only give the large sample methods, since these generalize to the case of censored data.

### *1.3.5 Point estimation of population percentiles*

For each positive fraction $p$, $\xi_p$ is called the population $100p$ percentile if

$$P(X \leq \xi_p) = p, \quad i.e.\ F(\xi_p) = p. \quad (1.74)$$

This percentile can also be defined as

$$S(\xi_p) = 1 - p. \quad (1.75)$$

It is easy to see that the population median is $\xi_{0.5}$. The above implicit definition can be reworded as

$$\xi_p = F^{-1}(p) = S^{-1}(1 - p).$$

The inverse function of $F$ is called the population *quantile function* and is denoted by $Q(.)$. In other words, for $0 < p < 1$,

$$Q(p) \equiv F^{-1}(p) = \xi_p.$$

In connection with the estimation of population percentiles, the inverse of the sample distribution function is useful. This function is denoted by $Q_n(p)$ and is called the *sample quantile function*. For each positive fraction $p$, it is defined as

$$Q_n(p) \equiv F_n^{-1}(p) = inf\{x : F_n(x) \geq p\}, \quad (1.76)$$

Copyrighted Material

where $F_n(.)$ is the sample distribution function. This definition of the inverse function is needed since $F_n$ is a step function. This simply means that $Q_n(p)$ is the smallest $x$-value such that $F_n(x)$ is not less than $p$ for the first time. Let $j = \lfloor np \rfloor$. If there are *no ties* in the sample, it is easy to see that for $0 < p < 1$,

$$Q_n(p) = \begin{cases} X_{(j)}, & \text{if } np = j; \\ X_{(j+1)}, & \text{if } np > j, \end{cases}$$

where $X_{(j)}$ is the $j$th order statistic. This definition results in one order statistic and it is generally used in asymptotic discussions. This $Q_n(.)$ function is used for estimating the percentiles. A point estimate of the $100p$ percentile $\xi_p$ is

$$\hat{\xi}_p = Q_n(p).$$

This estimate can also be expressed in terms of the sample survival function, $S_n(.)$. It is easy to see that

$$\begin{aligned} \hat{\xi}_p &= Q_n(p) \\ &= inf\{x : F_n(x) \geq p\} \\ &= inf\{x : 1 - S_n(x) \geq p\}. \end{aligned}$$

Finally, we have

$$\hat{\xi}_p = inf\{x : S_n(x) \leq (1 - p)\}. \tag{1.77}$$

Remark 1.4. The last expression can easily be applied to cases where the data contain some right-censored observations.

In our breakdown time example, discussed in Subsection 1.3.3, the estimate of the first quartile $\xi_{0.25}$, is $X_{(3)} = 18$ hours and the estimate of the median $\xi_{0.50}$ is $X_{(5)} = 24$ hours.

### *1.3.6 Confidence intervals for percentiles*

Suppose we want a $100(1-\alpha)\%$ confidence interval for the $100p$ percentile $\xi_p$. Let us consider the order statistics $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ of the random sample. (We are assuming that there are no ties.) These order statistics partition the real line into $(n + 1)$ intervals. We first compute the probability that $\xi_p$ belongs to the half open interval $[X_{(i)}, X_{(i+1)})$. We have

$$\begin{aligned} P(X_{(i)} \leq \xi_p < X_{(i+1)}) &= P(\textit{exactly } i \textit{ values are } \leq \xi_p) \\ &= \binom{n}{i} [F(\xi_p)]^i [1 - F(\xi_p)]^{n-i} \\ &= \binom{n}{i} p^i (1-p)^{n-i} \end{aligned}$$

Copyrighted Material

Now considering the union of such succesive intervals, we get the interval $[X_{(i)}, X_{(j)})$ and

$$P(X_{(i)} < \xi_p < X_{(j)}) = P(X_{(i)} \leq \xi_p < X_{(j)})$$
$$= \sum_{l=i}^{j-1} \binom{n}{l} p^l (1-p)^{n-l} \equiv C(i,j). \tag{1.78}$$

Thus the interval $(X_{(i)}, X_{(j)})$ is a confidence interval for $\xi_p$, with confidence coefficient $C(i,j)$. Generally, the confidence coefficient is chosen in advance. Thus we need to choose integers $i$ and $j$ such that $C(i,j)$ is at least $(1-\alpha)$, the chosen confidence coefficient. In other words, we choose integers $i$ and $j$ so as to satisfy the condition

$$P(i \leq Bin(n,p) \leq j-1) \geq 1-\alpha. \tag{1.79}$$

Now the interval $(X_{(i)}, X_{(j)})$ will be a $100(1-\alpha)$ confidence interval for $\xi_p$. It is obvious that more than one pair of integers $(i,j)$ will satisfy the condition (1.79). For some additional remarks about the choice of $i$ and $j$ see Appendix A1.

One choice of $i$ and $j$ (as given in Appendix A1) is that

$$P(Bin(n,p) < i) \leq (\alpha/2), \quad P(Bin(n,p) > j-1) \leq (\alpha/2).$$

A computer program has been developed for this purpose and is given in Appendix B1.

A method for determining a lower confidence bound is also given in Appendix A1.

Example 1.7. Suppose we want a 95% confidence interval for the first (lower) quartile, $\xi_{0.25}$. With $n = 10$, from the output of the computer program, we have $i = clower + 1 = 1$ and $j - 1 = cupper - 1 = 5$. Hence $j = 6$ and a 95% confidence interval is $(X_{(1)}, X_{(6)})$. For the data of Example 1.5, this interval is (4, 29).

Remark 1.5. In this discussion we assumed that there are no tied observations. If there are tied observations, we proceed as follows. For the integers $i$ and $j$ determined to satisfy (1.79), find the quantiles $\xi_p^L = Q_n(i/n)$ and $\xi_p^U = Q_n(j/n)$, where $Q_n(.)$ is the sample quantile function. The resulting confidence interval is $(\xi_p^L, \xi_p^U)$. Further details are available in Hutson (1999).

### *1.3.7 Kolmogorov's goodness-of-fit test*

Sometimes we want to test a simple null hypothesis about the population distribution function. In other words, the null hypothesis is

$$H_0 : F(x) = F_0(x), \tag{1.80}$$

Copyrighted Material

where $F_0$ is a completely specified cdf and the two-sided alternative is

$$H_A : F(x) \neq F_0(x), \textit{for some } x. \tag{1.81}$$

The test proposed by Kolmogorov tells us to evaluate the closeness of the sample distribution function $F_n(x)$ to the hypothesized cdf $F_0(x)$. The suggested closeness measure is

$$D_n = sup_x[\mid F_n(x) - F_0(x) \mid]. \tag{1.82}$$

This is the test statistic and an $\alpha$-level test

$$\textit{rejects } H_0 \textit{ if } D_n \geq C_{1-\alpha}. \tag{1.83}$$

It should be noted that the null distribution of the test statistic $D_n$ does not depend on $F_0(x)$. So this test is a distribution-free test. Birnbaum (1952) tabulated the distribution of $D_n$ and gave a table of the critical values for $\alpha = 0.05$ and 0.01. An extensive table of percentage points is contained in Miller (1956).

To implement the test, a convenient formula for computing the test statistic is needed. This expression for the statistic will enable us to infer that the null distribution of the statistic is independent of the distribution $F_0$. For simplicity, let us assume that there are no ties. We observe that the order statistics $(X_{(1)} < \cdots < X_{(n)})$ partition the real line into $(n+1)$ intervals and the sample distribution, $F_n$, is constant in each of these intervals. These $(n+1)$ intervals, which constitute a partition of the real line, are $I_0 = (-\infty, X_{(1)}), I_j = [X_{(j)}, X_{(j+1)})$, for $j = 1, \ldots, n-1$, and $I_n = [X_{(n)}, \infty)$. First we note that

$$D_n = max_j\{sup_{x \in I_j} |F_n(x) - F_0(x)|\}.$$

Next we calculate each of the supremums. It is easy to see that

$$sup_{x \in I_0} |F_n(x) - F_0(x)| = sup|0 - F_0(x)| = F_0(X_{(1)}),$$

and

$$sup_{x \in I_n} |F_n(x) - F_0(x)| = sup|1 - F_0(x)| = 1 - F_0(X_{(n)}).$$

For $j = 1, \ldots, n-1$, we have

$$sup_{x \in I_j} |F_n(x) - F_0(x)| = max\{(j/n) - F_0(X_{(j)}),\ F_0(X_{(j+1)}) - (j/n)\}.$$

Using the supremums in the $(n+1)$ intervals we have

$$D_n = max_{0 \leq j \leq n}[max\{(j/n) - F_0(X_{(j)}), F_0(X_{(j+1)}) - (j/n)\}]. \tag{1.84}$$

We note that $F_0(X_{(0)}) = 0$ and $F_0(X_{(n+1)}) = 1$. Under the null hypothesis the joint distribution of $(F_0(X_{(1)}), \ldots, F_0(X_{(n)}))$ is the same as the joint distribution of the order statistics of a sample of size $n$ from the uniform distribution on the interval $(0.1)$. Thus the statistic $D_n$ does not depend on $F_0$, which

Copyrighted Material

implies that $D_n$ is a distribution-free statistic. The above formula (1.84) for the statistic is equivalent to

$$D_n = max\{D_n^+, D_n^-\}, \tag{1.85}$$

where

$$D_n^+ = max_{0\le j\le n}[(j/n) - F_0(X_{(j)})], \tag{1.86}$$

and

$$D_n^- = max_{0\le j\le n}[F_0(X_{(j+1)}) - (j/n)]. \tag{1.87}$$

The expressions for the statistics $D_n^+$ and $D_n^-$ can be simplified as follows:

$$D_n^+ = max\left[0, max_{1\le j\le n}\left\{\frac{j}{n} - F_0(X_{(j)})\right\}\right],$$

and

$$\begin{aligned} D_n^- &= max\left[max_{1\le j\le n}\left\{F_0(X_{(j)}) - \frac{(j-1)}{n}\right\}, 0\right] \\ &= max\left[max_{1\le j\le n}\left\{F_0(X_{(j)}) - \frac{j}{n}\right\} + (1/n), 0\right]. \end{aligned}$$

It should be noted that these expressions are valid only for data sets with no ties.

Tests for one-sided alternatives

Even though these cases are of secondary importance, the test statistic $D_n$ turned out to be a function of the two statistics $D_n^+$ and $D_n^-$. These two statistics, $D_n^+$ and $D_n^-$ are useful for testing one-sided alternatives. For the alternative

$$H_+ : F(x) > F_0(x), \quad for\ some\ x, \tag{1.88}$$

the critical region is $D_n^+ \ge C_{1-\alpha}^+$ and for the alternative

$$H_- : F(x) < F_0(x), \quad for\ some\ x, \tag{1.89}$$

the critical region is $D_n^- \ge C_{1-\alpha}^-$.

Null distributions of $D_n^+$ and $D_n^-$

Birnbaum and Tingey (1951) derived the null distribution of $D_n^+$, and they showed that

$$P(D_n^+ \ge c|H_0) = c\sum_{j=0}^{J}\binom{n}{j}(1-c-(j/n))^{n-j}(c+(j/n))^{j-1} \equiv \pi(c), \tag{1.90}$$

where $J = \lfloor n(1-c)\rfloor$. So $C_{1-\alpha}^+$ is the solution of the equation

$$\pi(C_{1-\alpha}^+) = \alpha.$$

Copyrighted Material